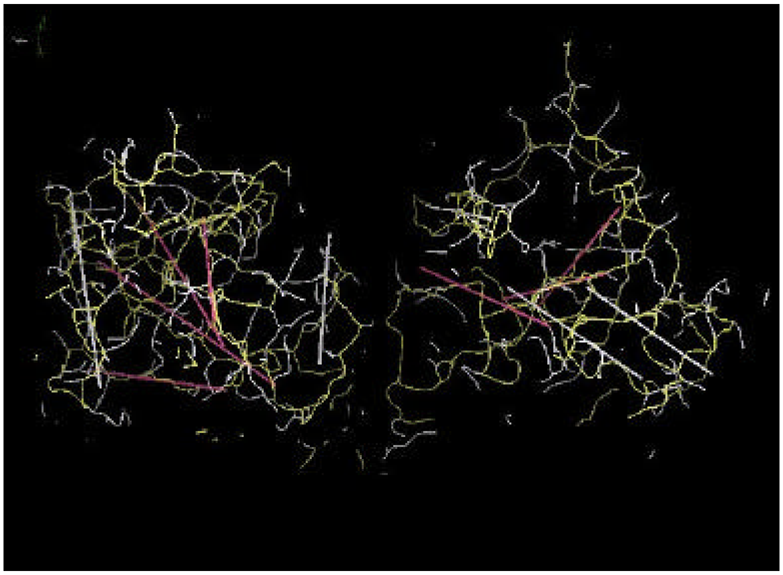
Molecular Modeling and Bioinformatics

Molecular modeling and simulation has become a fundamental technology for researchers engaged in structural biology for the characterization and design of proteins. Structural Biology researchers use a variety of software packages for structure determination and to understand the relationship between structure and function as they analyze and modify proteins and peptides. Protein modeling integrates structure generation, atomistic simulation, protein database searching and analysis, molecular visualization, and presentation. Understanding protein structure and function is essential to pharmaceutical and biotechnology companies and important in a range of other industries, including cosmetics, food, and agrochemicals.

X-ray Crystallography

Protein structure determination is a pre-requisite for structure-based research programs in the pharmaceutical and biotechnology industries. The computational approaches involved in protein crystallography include data processing and reduction, rapid model building, structure refinement, and graphical model manipulation. In today's competitive research environment, speed of results is of the utmost importance. Since X-ray crystallography has become a crucial step in the drug discovery process, crystallographers are under constant pressure to produce rapid results. Delays at the structure analysis step can lead to a bottleneck in the drug discovery pipeline. Recent developments in recombinant DNA techniques, crystallization protocols, X-ray data collection techniques and devices, and computing, have led to a substantial increase in the speed and number of protein structure determinations in modern crystallographic laboratories. However, there still remain a number of key stages in the crystallographic process, which limit the rate of structure determination. One of these is fitting electron density maps, either in the initial stages of tracing a chain to a new map, or in the manual rebuilding during refinement. This is a particularly onerous task, requiring many days and often weeks of working at a graphics terminal with maps and model. The figure below shows the results of an automated automated process of map interpretation, and the result of the analysis is overlayed on the bones.



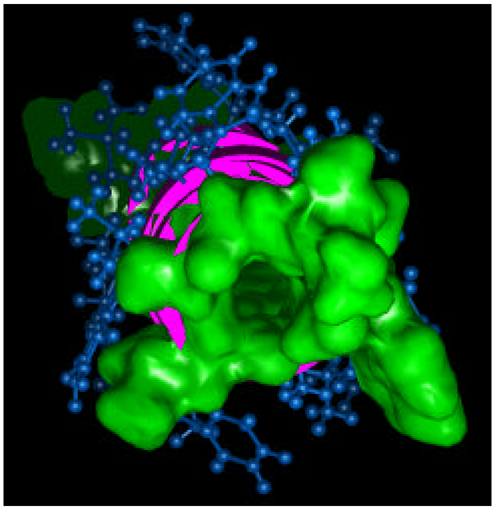
NMR

NMR is also a key component of structure-based research programs in the pharmaceutical and biotechnology industries. NMR as a method for biomolecular structure determination includes spectral data processing through to structure refinement and evaluation. The process of data analysis and structure determination has been streamlined by employing an integrated set of modeling, assignment, and analysis tools.

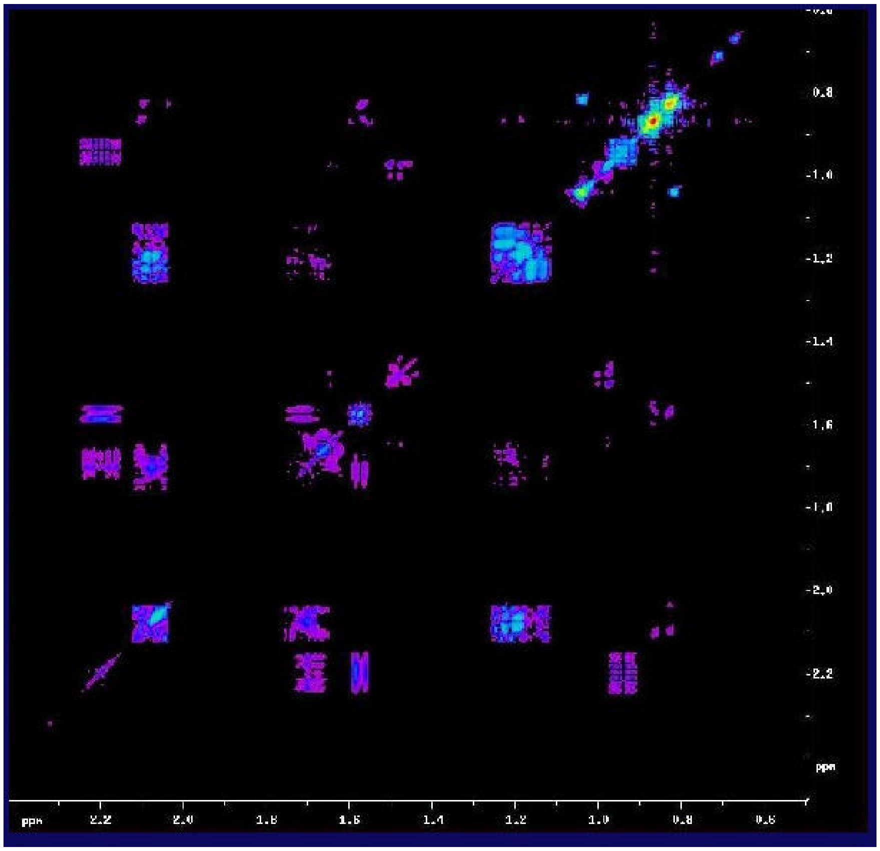
Gramicidin is a polypeptide antibiotic, which forms cation-specific channels in membranes. The conductance properties of this molecule have been extensively characterized, thus making it a good candidate for elucidating structure/function relationships for the process of ion transport. Using nuclear magnetic resonance, circular dichroism, and vibrational spectroscopy, X-ray crystallography and molecular graphics techniques, the orientation, conformational flexibility, and effects of ion binding, lipid structure and phase state on the structure of this molecule have been demonstrated, the different folding motifs it adopts in membranes and in solution. Chemical modification studies and de novo synthesis have produced alterations in conductance properties that can now be understood on a molecular level. Studies by Bonnie Wallace's group have provided the first high resolution view of an ion channel in complex with a ligand (caesium) and insight into its mechanism of action, which forms the basis for molecular designs of new synthetic channels.

Current work includes studies on the nature of ion binding using crystallography to determine the altered types of interactions of the polypeptide backbone with the cation as it traverses the pore and CD spectroscopy to determine binding constants for a series of monovalent cations, as well as the structure of the blocked pore by 2-D NMR. In addition, crystals of a complex between gramicidin and lipid molecules have recently been prepared, and structure determination is in progress. Other studies are involved in elucidating the process of insertion of this polypeptide into the hydrophobic environment of the lipid bilayer, and examining the process of unfolding and refolding the molecule as it converts between the channel and pore conformations.

In the graphic below, we can look down the channel of the antibiotic gramicidin, which is used to transport potassium ions across mitochondrial and bacterial membranes. It has been rendered as a molecular surface with a ribbon.



Shown below is a 2D NMR (gradient HCOSY) spectrum of gramicidin.



Protein Engineering

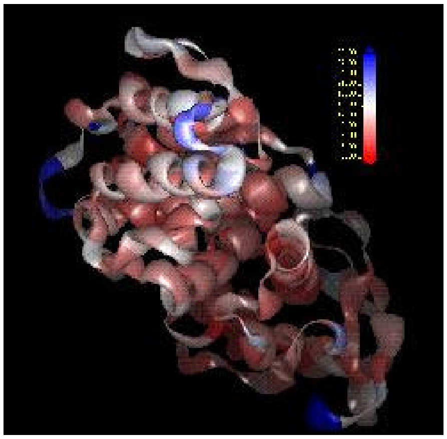
Understanding protein structure and function is essential to the development of new drugs. Researchers engaged in the characterization and design of proteins need to understand this relationship as they analyze and modify proteins and peptides. Searching sequence and structural protein databases for similar proteins, predicting protein structure, and modeling these structures, are all important steps in a comprehensive drug development program.

Life Science Simulation

Molecules are not static structures. They move, vibrate, and interact with other molecules and their environment. Understanding these movements and interactions is essential for the complete understanding of structure-function relationships including many aspects of drug design and intermolecular interactions.

Molecular dynamics simulations include model building of small molecules and biomolecules, graphical model manipulation, energy minimization, graphical trajectory display, and data analysis. The heart of molecular simulations is the forcefield engine, the workhorse for the calculations. In molecular mechanics and dynamics calculations, the forcefield determines how the system under study will behave and if your results will be valid. Therefore, it is absolutely essential that you use the best set of parameters developed for the system of interest. Data analysis is critical for getting meaningful answers.

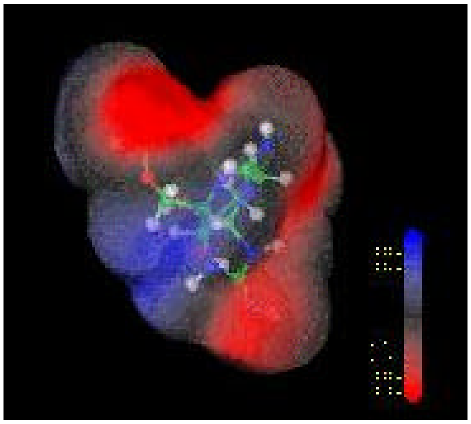
One example of a forcefield is hydrophobicity. In the graphic below, we have used a ribbon to show how this property varies along the backbone of the protein. The hydrophobicity of each residue is shown by the ribbon width and the temperature factor of each residue's carbon alpha is shown by color according to the spectrum.



Molecular dynamics simulations on large systems also involve docking of ligands in protein receptor sites, conformational changes such as HIV flap movement and subunit shifts in hemoglobin between oxy and deoxy states, protein folding and diffusional conformational changes.

A limitation in molecular dynamics simulations is that your highest frequency vibrations limit the timestep size in the system under study. For biological molecules, the highest frequency modes are found in C-H bond stretching and thus limit timesteps to 1-2 femtoseconds. Unfortunately, many of the interesting molecular motions have lower frequency modes. Thus to study them you must run longer simulations.

Another force field component is the electrostatic potential. In the example on the next page, saxitoxin, a neurotoxin, has been rendered with an electron density contour surface. The surface is color-coded by electrostatic potential, using graduated transparency. The spectrum shows that negative and positive regions are opaque red and blue, respectively, and neutral regions are transparent. A ball-and-stick rendering of the structure is seen inside the electrostatic potential surface.



The solvent environment of biological systems can play a major role in their activity and interaction with other molecules. However, the computational expense of including explicit solvent into calculations is often prohibitive. Also, the determination of electrostatic potentials and solvation energies can be a daunting task without the proper tools.

A Brief Description Traditional Homology Methods

Homology based methods have been useful tools for protein structure determination since the early 1970's. Traditional methods involve the following steps:

- identification of a sequence homology between a protein of known structure and an "unknown" protein.
- superimposition, based on alignment of the "known" structures.
- alignment of the unknown's sequence to the sequences of the known structures
- copying of stretches of the known structures to the unknown, based on the alignment.
- patching of gaps in the unknown by fragment searches or template generation.
- refinement by molecular mechanics energy minimization, molecular dynamics, special loop searching algorithms, or Monte Carlo methods.

With this method protein structures can be determined from homology to about an RMSD of 1.0 to 2.0 angstroms for 80% of residues where there is 30-40% sequence identity.

Determination of the 3D structure of Sea Raven Antifreeze Protein

In this example an alignment is generated and then modeling methods are used to generate five conformers.

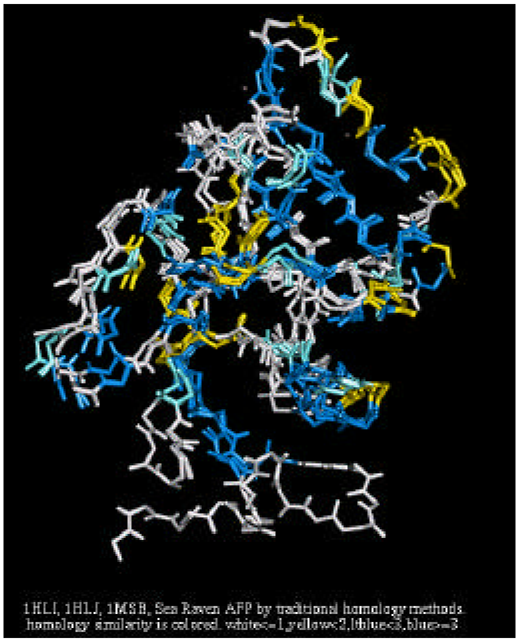
- Sequence Creation
  The "unknown" sequence is read from the EMBL/Swissprotein database. Other input possibilities include FASTA, GCG, NBRF-PIR, PDB

- Identification of Homologous Sequences
  Typically, the FASTA method of searching a protein database is used to identify homologous sequences. In this case, the FASTA search identifies three structures in the PDB (2MSB and 1HLI and 1HLJ) that have at least marginal homology to Sea Raven AFP. These PDB files are downloaded and inspected. Both 1HLI and 1HLJ are computationally derived structures with about 27% and 31% identity in 105 residues. 2MSB is about 14% identical to Sea Raven, however the key cysteine residues are all in regions that are conserved.

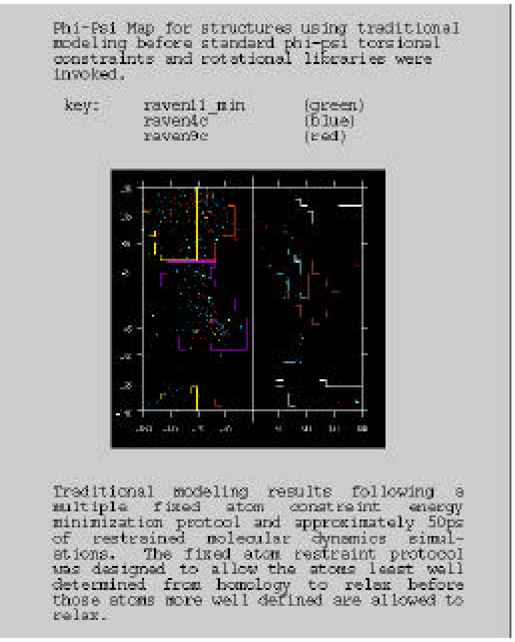
- Creation of an Alignment
  The alignment is generated using standard alignment tools. Modifications are made using alignment editing tools, based on the previous FASTA sequence-only alignments of the antifreeze proteins and type C lectins which have high homology to AFP, but for which structures are not known. If no alignment is given, an alignment can be generated automatically. There is no homology based information known for residues 1-15, but there is a cystine linkage known from experiment between residues 7 and 8.

- Traditional Homology Modeling
  Large gaps can be filled by fragment libraries and loop searching. A refinement protocol of restrained energy minimization (four stages with 2000 steps of energy minimization per stage) followed by 50 ps of molecular dynamics annealing is used. This might take as long as eight hours per structure on an SGI R4000. The output pdb files can be read into a molecular viewing program, overlayed and ready for analysis.

Shown below are figures of the results of this traditional homology modeling, and the Ramachandran phi/psi plot of the backbone angles.



1HLI, 1HLJ, 1MSB, Sea Raven AFP by traditional homology methods. homology similarity is colored. white<=1,yellow<2,ltblue<3,blue>=3



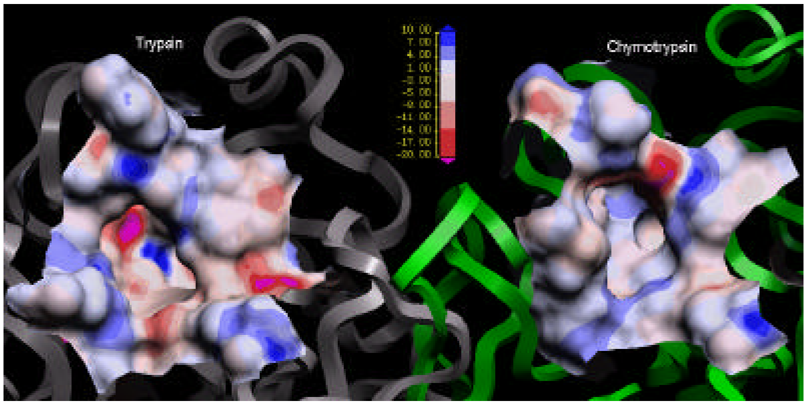
Phi-Psi Map for structures using traditional modeling before standard phi-psi torsional constraints and rotational libraries were invoked.

key:		
	raven11_min	(green)
	raven4c	(blue)
	raven9c	(red)

Traditional modeling results following a multiple fixed atom constraint energy minimization protocol and approximately 50ps of restrained molecular dynamics simulations. The fixed atom restraint protocol was designed to allow the atoms least well determined from homology to relax before those atoms more well defined are allowed to relax.

Investigating Protein Activity Using Molecular Modeling Techniques

Visualizing properties mapped onto molecular surfaces can be used to rationalize differences in the activity of proteins. This information can be used in drug design and protein engineering. The graphic below shows the electrostatic potential mapped onto the molecular surfaces of the proteins trypsin and chymotrypsin.



The molecular surfaces of the two proteins show active sites of remarkably similar structure: both have a "pocket" which is about the right size to accommodate an amino acid side chain. However, while both are proteases, they differ in their specificity. Trypsin specifically breaks peptide bonds to lysine and arginine, while chymotrypsin is specific for phenylalanine, tyrosine, and tryptophan. Since the pockets have similar shape, what accounts for the differing specificity?

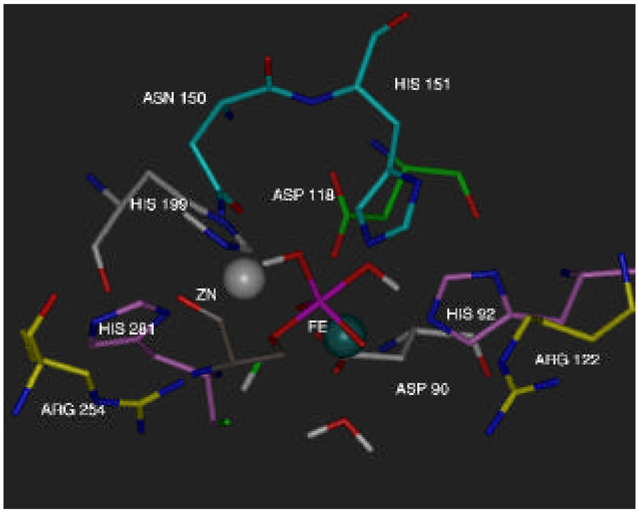
The difference can be accounted for by examining the electrostatic potentials within the pocket. The technique used for display is to color code the molecular surface according to the electrostatic potential at each point on the surface. The relationship between a color and an electrostatic potential value is shown on the accompanying spectrum. Using this technique the reason behind the differing specificities is clear: trypsin has a negatively charged region at the bottom of the pocket which chymotrypsin lacks. Thus, trypsin is specific to residues with positively charged side chains and chymotrypsin is specific to residues with large hydrophobic side chains.

Structure-Based Drug Design

Molecular graphics, the indispensable tool of the modern drug designer, has undergone tremendous enhancements in recent years. No longer simply a sophisticated visualization aid, molecular graphics has evolved into the platform for advanced calculations aimed at exploring the atomic properties of molecules from simple chemicals to complex proteins. Simulations of the chemical and physical behavior of multi-component systems have allowed, up to now, rare studies such as protein-ligand binding and energetics to become commonplace. Ready access to this new technology has opened opportunities for the development of powerful new strategies for drug design.

Understanding molecular recognition is central to the rational design of small molecule and protein pharmaceuticals. The shape and electrostatic properties of a molecule are two of the major determinants in molecular interactions. To elucidate the 3D structure or to view relationships between properties, fields and structures, it is crucial to have the ability to examine the structure in a variety of ways.

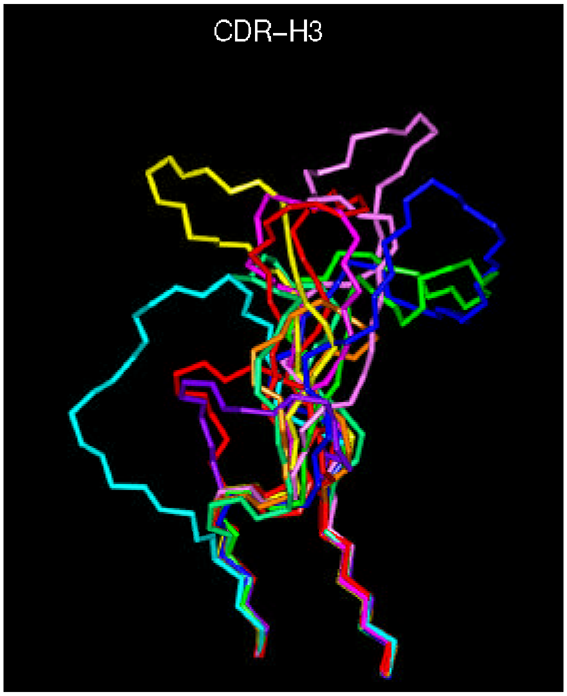
Agouron Pharmaceuticals, Inc. http://www.agouron.com has teamed up with Molecular Simulations Inc, http://www.msi.com to provide Agouron's structure-based drug designers the most advanced suite of computational tools available. These new methods are being applied to the discovery of drugs for viral infections, cancer and immunological diseases. One example of the application of computational simulations is Agouron's research into the mechanism of inhibition of calcineurin by the potent immunosuppressants FK506 and cyclosporin. To understand the chemical catalysis carried out by calcineurin, simulations of the transition state are being used to reveal the role of various active site residues and the two metal ions in the catalytic site. They have found that primarily the electrostatic field created by the active site residues stabilizes the catalyzed reaction's transition state. The active site residues and metal ions within the calcineurin catalytic site are shown in the figure on the next page.



Structural Classification of CDR-H3 in Antibodies

Scientists at the Biomolecular Engineering Research Institute (BERI) have determined crystal structures of segments of the third complementarity determining region of the antibody heavy chain (CDR-H3). Large variations in the lengths and amino acid sequences of the third complementarity determining region of the antibody heavy chain (CDR-H3) have made it difficult to establish a relationship between these sequences and their tertiary structures. This is in contrast to the other CDRs, which have been classified by their canonical structures. If however, CDR-H3 local structures were classified for some protein segments, and relationships were determined between amino acid sequences and their conformations for just a few specific sequences with particular features, such limited information would still be very helpful to build reliable three dimensional (3D) models of antibodies from amino acid sequences only.

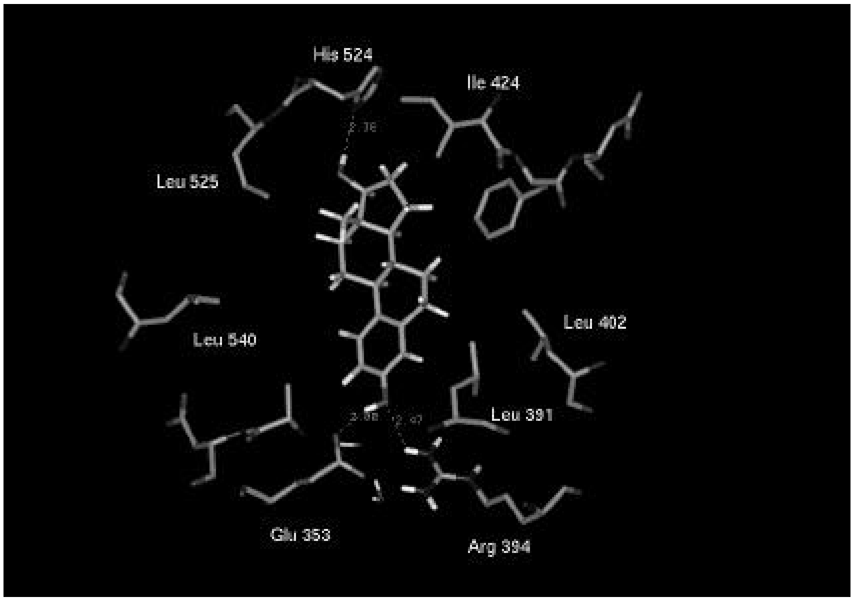
From the crystal structures shown at the right the BERI group has derived several novel rules, which could partly predict the CDR-H3 conformation from only sequence information (Shirai, H., Kidera, A. and Nakamura, H. FEBS Lett. 399, 1-8, 1996). Since these rules are physically reasonable, they are expected to be applicable to structural modeling and design of antibodies.



Modeling Of the Estrogen-Receptor Ligand Binding Domain

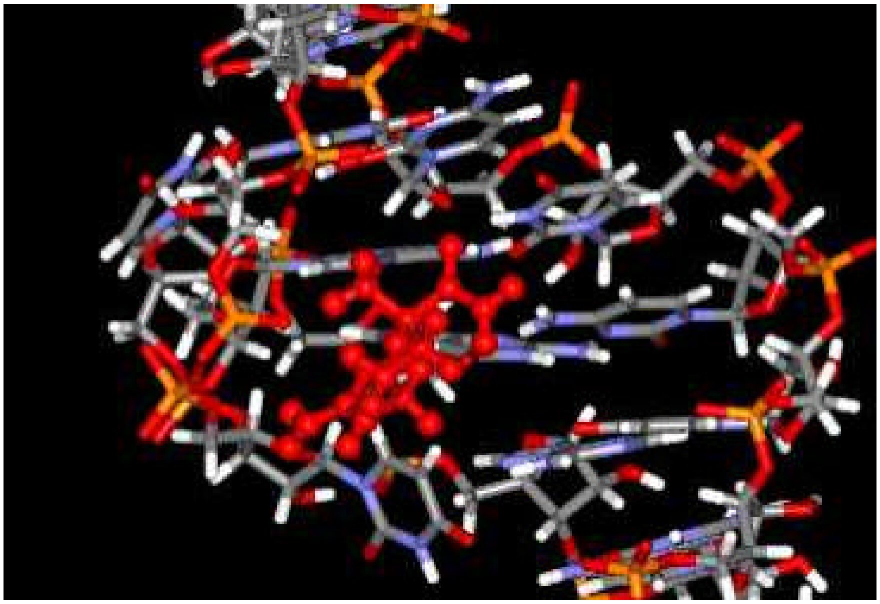
The estrogen receptor (ER) plays a crucial role in many processes such as the control of reproduction and the development of secondary sexual characteristics. The ER belongs to the superfamily of nuclear receptors (NRs), which includes the steroid, thyroid, and retinoic receptors. NRs are ligand activated transcription factors and modulate the expression of specific genes. All NRs consist of five to six domains, one of which is the ligand binding domain (LBD). Crystal structures have been determined for the LBDs of ER, the thyroid hormone receptor as well as the retinoic receptors RARgamma and RXRalpha. Only the co-ordinates of the latter two are deposited in the Brookhaven Protein Databank (PDB entries: 1LBD, 2LBD).

A three-dimensional model of the human ER LBD was generated by researchers at Schering AG on the basis of the human RARgamma LBD. The sequence identity between the LBDs of the ER and RARgamma receptor is only 21%, nevertheless it is generally assumed that all NR LBDs share the same fold. The quality of the model was tested against mutants, which affect the binding properties. A thorough analysis of all published mutants elucidated the effect of the mutations. 45 out of 48 mutants can be explained satisfactorily on the basis of the model. After that, the natural ligand estradiol was docked into the binding pocket to probe its interactions with the protein. Energy minimizations and molecular dynamics calculations were performed for various ligand orientations. The analysis revealed two favorite estradiol orientations in the binding niche of the ER LBD forming hydrogen bonds with Arg394, Glu353 and His524 (shown on the next page). Both models explain largely the binding affinities of more distantly related ligands. The crystal structure of the ER LBD in complex with estradiol was later published (Brzozowski et al. Nature **389**, 753-758, **1997**), and one of the models is largely in agreement with the experimentally determined structure.



Ligand Docking

If a complete ligand is known and a binding is suspected, you can employ docking tools to identify possible orientations of the ligand to the receptor. With these tools you can identify key interactions and possible ligand modifications to either enhance ligand binding or develop a novel compound. Molecules can be fit into the active site of a receptor by matching complementary polar and hydrophobic groups. An empirical scoring function is used to prioritize the hits. Modifications may then be suggested to increase the binding affinity between an existing ligand and the receptor. Combined Monte Carlo and simulated annealing protocols may be used to identify possible orientations of the ligand in the receptor site. In the image below, we see the orientation of an arginimide residue in HIV TAR RNA receptor. This area is recognized by the Tat protein through a series of arginine residues. Only one is required for specific binding of small peptides to this region.



Bioinformatics

The essence of bioinformatics is extracting new knowledge from existing collections of biological data. Therefore, bioinformatics tools are only as powerful as the data they access. Protein bioinformatics involves the annotation of a protein sequence at the primary, secondary and tertiary structure levels to maximize the information and knowledge of that sequence. The aim is to assign a function to a protein sequence and to characterize its interactions, activities and role in normal and diseased states

A revolution is occurring in molecular biology. As the Human Genome Project fulfills its promise to provide sequences for all human genes, molecular biologists are increasing their reliance on computational methods. Newly discovered genes can often be characterized easily through their relationship with other known genes by comparing the sequence or structure of the new gene with that of all known genes. Computational biology is proving to be a far simpler and quicker way to begin to investigate genes than experimentation.

With these new opportunities comes a burden, however. The personal computers that molecular biologists use today are no longer capable of keeping up with the demands. The rapid advancement in bioinformatics technology and the multidisciplinary nature of many tools makes thorough and accurate gene analysis formidable. Powerful bioinformatics methods exact a high computational price. This combined with the exponential growth in the size of genomic databases makes it impossible to be effective with only a desktop computer.

Until recently, the solution to this dilemma has been to use workstations and terminal emulators to run jobs remotely. Most molecular biologists are not interested in learning UNIX or command-line driven programs, and as a result less computational analysis takes place. The World Wide Web (WWW) has provided an interesting solution to this problem by presenting users with the functionality of UNIX programs from the environment of a web page. Numerous programs have been adapted in this way, and it is proving to be an effective approach.

Finally, there are many different types of tasks to be performed in the process of doing gene research. Thus, there are many different software tools available for use. Because most of the available tools have been created independently, they are often very different from each other in form and substance. Users must therefore learn multiple interfaces and manually transfer data between applications. As this field matures, there is a strong desire for comprehensive integrated tools. Rather than negotiating a way through a series of focused niche products, each of which requires additional training and reformatting of data, users are searching for a seamless path for data analysis. The use of WWW based interfaces goes a long way toward that goal, and in the process opens up the world's resources.

The study of genes and proteins has been greatly enhanced with the addition of sequence similarity bioinformatics tools. However, the emergence of new bioinformatics technology based on protein structure promises to quicken the transformation of the field.

The first indications of the function of a newly sequenced gene are often due to similarities observed between its protein sequence and that of another protein of known function. This conclusion relies on the observation that proteins similar in sequence are also similar in function. However, the function of a protein is often determined more directly by its three-dimensional (3D) structure than by its amino acid sequence. Thus 3D bioinformatics methods, including 3D protein database searching and analysis techniques, offer several advantages that are not currently possible by using traditional sequence matching methods in isolation.

Threading technology can often find protein homologs that sequence similarity searching fails to detect. Threading is effective because its comparisons involve structural components of proteins, which may be more conserved than the underlying amino acid sequence. Evolution ultimately conserves protein function, which is based on 3D structure, which is based on 2° structure, based on protein sequence based on DNA sequence

3D model structures can often be predicted for a new protein sequence. A far greater understanding of a protein can be achieved with a 3D structure than with a linear sequence alone. The ability to view a folded protein in three dimensions can make interesting features of the protein clear. For example, when functional mutations are mapped onto the model, clustering of mutations in 3D space will often highlight critical structural features.

Mechanisms for the ligand specificity of a family of proteins can be investigated by comparing subtle differences in the 3D structures of the family members. Visual inspection of the size and conformation of the active site or surface electrostatics can prove quite illuminating.

One of the greatest driving forces pulling on genomic data is the desire to find therapeutic drugs. 3D bioinformatics brings genomic data one step closer to this process. Current technology can discriminate genetic targets by molecular docking potential, for example. It is no longer possible to fully characterize a gene and its protein product without relying on the technologies of 3D bioinformatics. Today, dependence on sequence based methods alone may leave many valuable insights unrevealed.

The number of known protein structures is growing much more slowly than the number of protein sequences. As a result the need for structure determination methods from homology is acute. Protein structures generated from homology are an important part of any drug discovery strategy. The protein models built using 3D bioinformatics can also be used as starting points for X-ray or NMR structure refinement. Typical problems include modeling renin or HIV protease to known structures of aspartyl proteases; modeling TPA to known structures of serine proteases; modeling new or modified immunoglobulins to known structures of the family; generating starting models for NMR structure determination; and generating starting models for molecular replacement and model fitting in X-ray crystallography of homologous or mutant proteins.